Predict Customer Personality to boost marketing campaign by using Machine Learning

Created by:
St S Bintang Pratama Dumatubun
nivandumatubun30@gmail.com
Nivan Dumatubun

Data Enthusiast. I have an interest in data and technology, i'm also pursuing experience and knowledge in technology through bootcamp and projects.

Supported by:
Rakamin Academy
Career Acceleration School
www.rakamin.com

Overview

A company can develop rapidly when it knows its customers' personality behavior, so that it can provide better services and benefits to customers who have the potential to become loyal customers. By processing historical marketing campaign data to improve performance and target the right customers so they can make transactions on the company's platform, from this data insight our focus is to create a cluster prediction model to make it easier for companies to make decisions.

Feature Engineering

Pada tahap *feature engineering*, dilakukan beberapa penambahan fitur diantaranya yaitu :

1. Membuat kategori umur dimana kategori umur didasarkan pada rentang usia customer dengan kategori :
   - Child
   - Teen
   - Young Adults
   - Middle Aged Adults
   - Older Adults

2. Membuat total campaign yang diterima oleh customer dengan menjumlahkan kolom accepted campaign 1 hingga 5

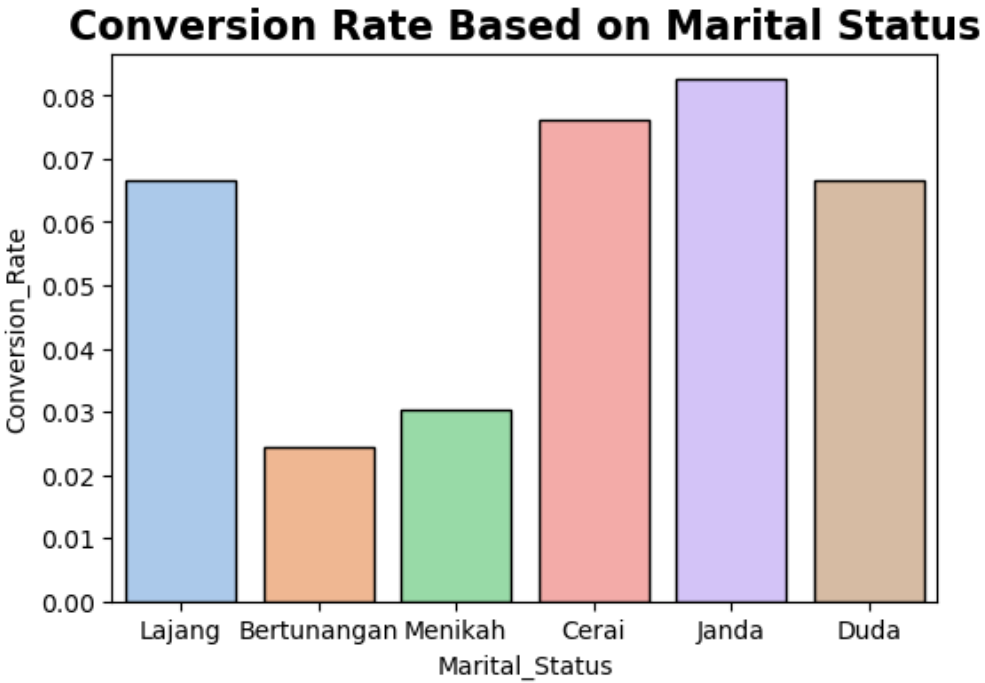
3. Conversion Rate, dibuat dengan membandingkan jumlah respon customer dengan jumlah kunjungan situs internet sehingga diperoleh conversion rate

4. Total Children dan is_parents, dibuat dengan menjumlahkan kolom Kidhome dengan Teenhome

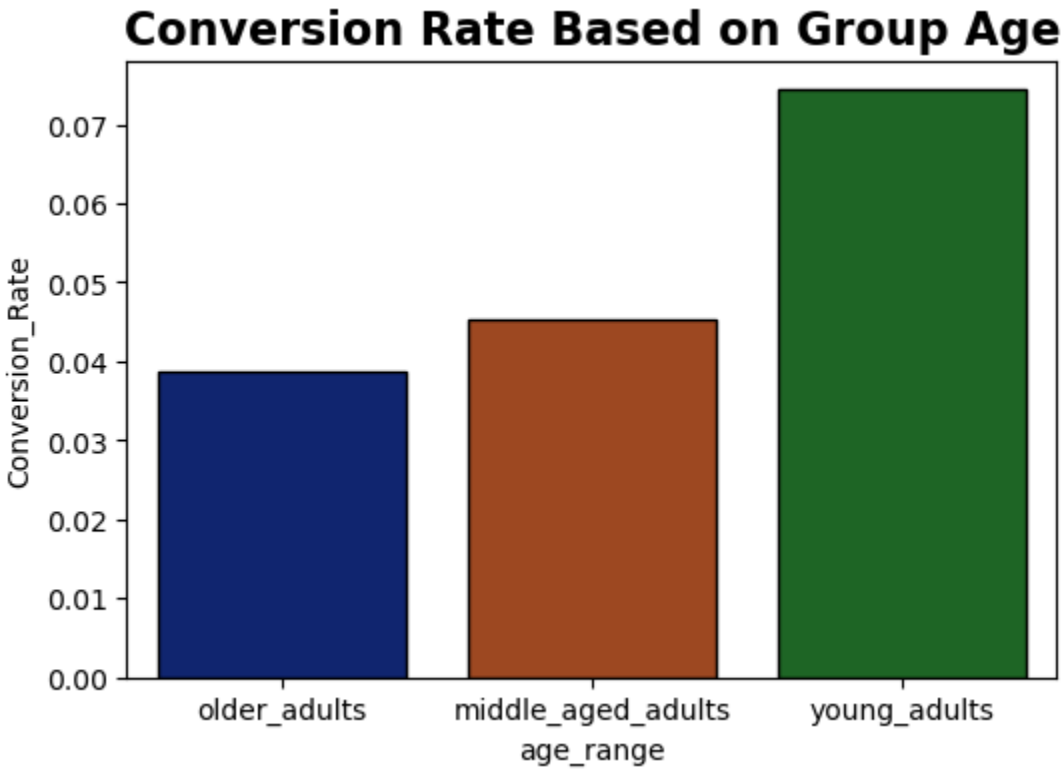
5. Total Transaction, dibuat dengan menjumlahkan NumDealsPurchases, NumWebPurchases, NumCatalogPurchases, dan NumStorePurchases

6. Total Accepted Campaign, dibuat dengan menjumlahkan Accepted Campaign #1 hingga #5

Insight



Setelah dibuat grafik perbandingan antara *conversion rate* dengan *marital status*, dapat dilihat bahwa conversion rate tertinggi dimiliki oleh customer dengan status **janda** sedangkan yang terendah **bertunangan**. Mayoritas customer yang memiliki conversion rate > 0.5 yaitu berstatus **Lajang**, **Cerai**, **Janda**, dan **Duda**. Sehingga mungkin dapat campaign dapat ditargetkan kepada customer dengan status tersebut diatas.



Mayoritas customer dengan conversion rate tinggi adalah seorang young_adults yang berusia 18 – 36 tahun. Sedangkan yang terendah yaitu older_adults yang berumur > 55 tahun. Sehingga mungkin campaign dapat ditargetkan kepada customer young_adults yang mana memiliki conversion rate yang cukup tinggi.

Data Cleaning

2. Data Cleaning

```
[13]: df1 = df.copy()
```

2.1 Handling Missing Value

```
[14]: df1['Income'].fillna(df['Income'].median(), inplace=True)

[15]: df1.isna().sum()

[15]: ID                    0
      Year_Birth            0
      Education             0
      Marital_Status        0
      Income                0
      Kidhome               0
      Teenhome              0
      Dt_Customer           0
      Recency               0
      MntCoke               0
      MntFruits             0
      MntMeatProducts       0
      MntFishProducts       0
      MntSweetProducts      0
      MntGoldProds          0
      NumDealsPurchases     0
      NumWebPurchases       0
      NumCatalogPurchases   0
      NumStorePurchases     0
      NumWebVisitsMonth     0
      AcceptedCmp3          0
      AcceptedCmp4          0
      AcceptedCmp5          0
      AcceptedCmp1          0
      AcceptedCmp2          0
      Complain              0
      Z_CostContact         0
      Z_Revenue             0
      Response              0
      dtype: int64
```

Terdapat *missing value* pada kolom *Income* sehingga dilakukan pengisian baris kosong tersebut dengan nilai Tengah (median).

```
df.duplicated().sum()

0
```

Setelah dilakukan pemeriksaan terhadap data duplikat, tidak terdapat baris duplitkat pada dataframe ini.

Feature Encoding

Label Encoding

```
df1['Education'].value_counts()
```

```
S1     1127
S3      486
S2      370
D3      203
SMA      54
Name: Education, dtype: int64
```

```
edu_mapping = {'SMA' : 1,
               'D3' : 2,
               'S1' : 3,
               'S2' : 4,
               'S3' : 5}
df1['Education'] = df1['Education'].map(edu_mapping)
```

```
df1['Education'].value_counts()
```

```
3    1127
5     486
4     370
2     203
1      54
Name: Education, dtype: int64
```

```
df1.sample(3)
```

	Year_Birth	Education	Income
204	1965	5	40637000.0
533	1962	3	65316000.0
1521	1971	3	69930000.0

Label encoding dilakukan pada kolom *Education* karena kolom tersebut mengandung value yang bertipe ordinal sehingga digunakan label encoding sebagai encoder kolom tersebut. Dapat dilihat gambar disamping menunjukkan bahwa value pada kolom *Education* sudah menjadi angka.

Feature Encoding

One Hot Encoding

```
onehot = pd.get_dummies(df1['Marital_Status'],prefix = 'status')
df2 = df2.join(onehot)
```

```
df2.drop('Marital_Status',axis=1,inplace=True)
```

```
onehot1 = pd.get_dummies(df2['age_range'])
df2 = df2.join(onehot1)
df2.drop('age_range',axis=1,inplace=True)
```

status_Bertunangan	status_Cerai	status_Duda	status_Janda	status_Lajang	status_Menikah	middle_aged_adults	older_adults	young_adults
0	0	0	0	1	0	0	1	0
0	0	0	0	0	1	1	0	0

One Hot Encoding dilakukan pada kolom *Marital Status* dan *Group Age*. Hal ini dikarenakan value pada kedua kolom tersebut bukan merupakan tipe data ordinal atau yang dapat diurutkan. Gambar disamping menunjukkan pseudocode dari one hot encoding kedua kolom dan kolom baru yang tercipta dari One Hot Encoding yang sudah dilakukan.

Feature Standardization

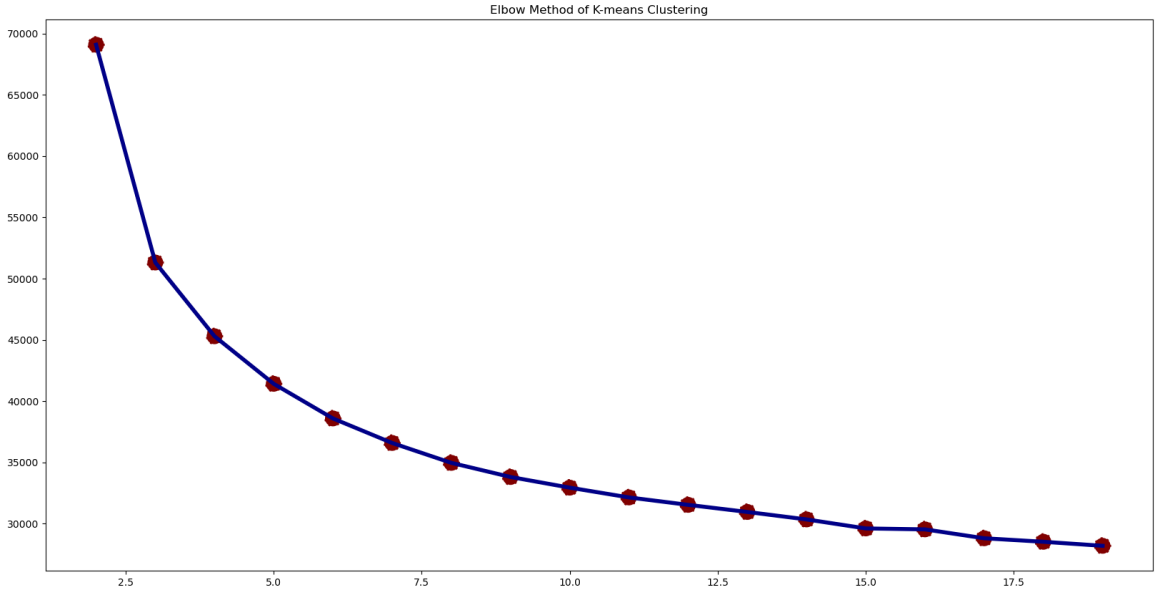
```
df_std = df2.copy()
```

```
from sklearn.preprocessing import StandardScaler
ss = StandardScaler()

for col in num:
    df_std[col] = ss.fit_transform(df_std[[col]])

display(df_std.shape, df_std.head(3))
```

```
(2216, 42)
```

Gambar disamping merupakan potongan *code* dari proses standardisasi yang telah dilakukan. Dengan demikian proses standardisasi sudah berhasil dilakukan.

Elbow Method

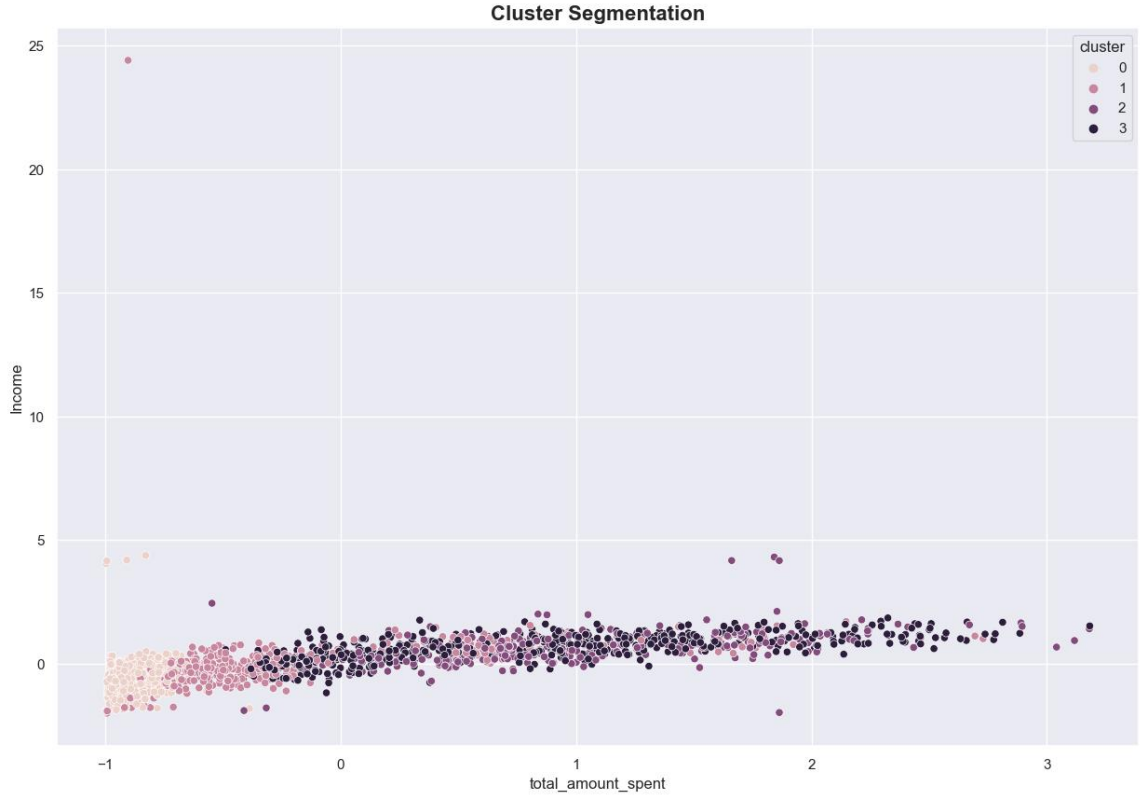



```
(pd.Series(inertia) - pd.Series(inertia).shift(-1)) / pd.Series(inertia) * 100
```

```
0     25.731437
1     11.729431
2      8.570029
3      6.825886
4      5.187254
5      4.439742
6      3.298184
7      2.575044
8      2.423072
9      1.886631
10     1.839767
11     1.968628
12     2.420064
13     0.262929
14     2.458262
15     0.993578
16     1.153455
17          NaN
```

Terlampir diatas merupakan *Elbow Graph* yang dihasilkan dari elbow method. Berdasarkan elbow method yang telah dilakukan, dapat dilihat bahwa cluster yang optimal yaitu 4 cluster. Hal ini didukung oleh persentase inertia yang berkurang sebanyak 25.7% pada penambahan cluster dari 1 menjadi 2, dan pengurangan 11.7% inertia pada penambahan cluster dari 2 menjadi 3 dan 8.5% pada penambahan cluster 3 menjadi 4 sehingga yang dipilih 4 cluster.

Clustering

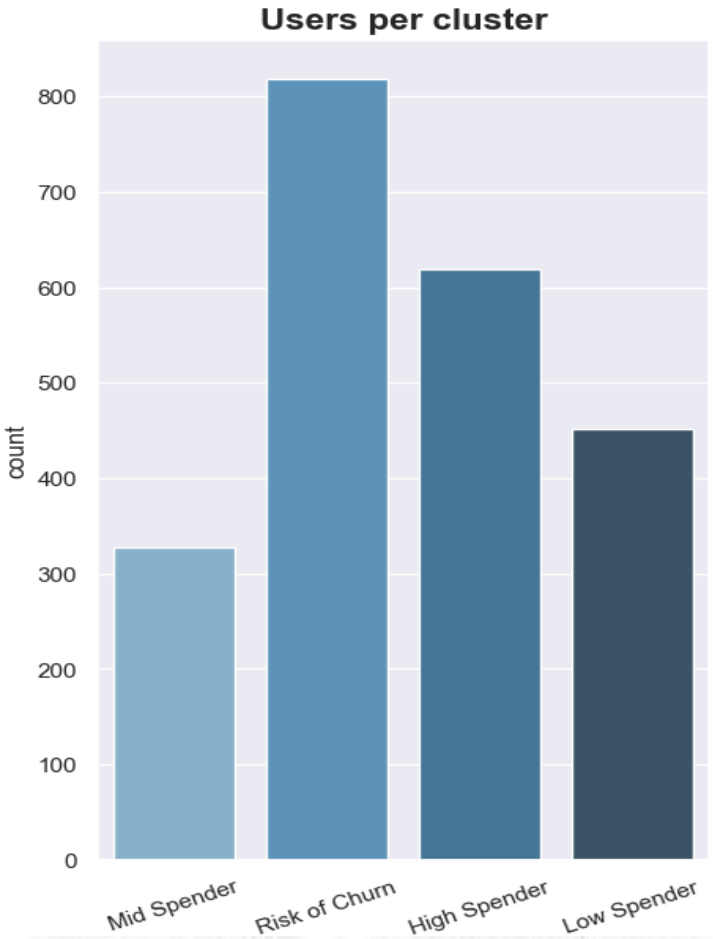


Berikut adalah *scatterplot* Cluster Segmentation. Dapat dilihat bahwa cluster 3 (hitam) merupakan kelompok dengan total pengeluaran tertinggi yang berbanding lurus dengan pemasukan yang juga tinggi.

Cluster dibagi menjadi 4 kelompok yaitu :

- Cluster 0 (**Risk of Churn**)
- Cluster 1 (**Low Spender**)
- Cluster 2 (**Mid Spender**)
- Cluster 3 (**High Spender**)

Distribusi Cluster



Berikut merupakan distribusi dari masing-masing cluster. Berdasarkan pengujian yang telah dilakukan, diperoleh jumlah masing-masing cluster yaitu sebanyak :

- Risk of Churn = 818 user
- Low Spender = 452 user
- Mid Spender = 327 user
- High Spender = 619 user

Mid Spender merupakan kelompok dengan jumlah user paling sedikit dan jumlah user terbanyak dimiliki oleh kelompok high spender.

Cluster Description

1. **Risk of Churn**:
- Kelompok ini adalah kelompok dengan jumlah user terbesar kurang lebih 800 orang yang di dominasi oleh middle_aged_adults (36-55 tahun), yang dominan telah menikah dan mempunyai 1 anak.
- Dari segi pendapatan dan pengeluaran, kelompok ini mempunyai pendapatan dan pengeluaran paling kecil di setiap bulannya, yang masing-masing sebesar IDR 33.4 Juta untuk total pendapatan setahun, dan IDR 53K untuk pengeluaran dalam setahun
- Walaupun demikian, kelompok ini menduduki peringkat pertama untuk kelompok yang paling sering mengunjungi web dengan median total kunjungan 7 kali dalam sebulan, walaupun demikian, mereka masih jarang untuk bertransaksi dan bahkan menggunakan promo pada transaksinya. Selain itu juga kelompok ini merupakan kelompok yang paling jarang atau hampir tidak pernah menerima *campaign*.

2. **Low Spender**:
- Kelompok ini didominasi oleh older_adults (>55 tahun) dan middle_aged_adults (36-55 tahun), yang dominan telah menikah dan mempunyai 1 anak
- Kelompok ini mengunjungi website cukup sering, kedua tersering setelah Cluster 0, dengan median sebanyak **6 kali dalam sebulan**, walaupun demikian, kelompok ini cukup jarang menerima *campaign*.
- Namun, kelompok ini mempunyai total pendapatan dan pengeluaran terkecil kedua dibandingkan Kelompok lainnya, yang masing-masing sebesar IDR 49 Juta untuk total pendapatan setahun, dan IDR 319K untuk pengeluaran dalam setahun

3. **Mid Spender**:
- Kelompok ini didominasi oleh older_adults (>55 tahun) dan middle_aged_adults (36-55 tahun), yang dominan telah menikah dan mempunyai 0-1 anak dan merupakan jumlah user paling sedikit
- Kelompok ini mempunyai total pendapatan dan pengeluaran terbesar kedua dibandingkan Kelompok lainnya, yang masing-masing sebesar IDR 67 Juta untuk total pendapatan setahun, dan IDR 1.1 Juta untuk pengeluaran dalam setahun
- Walaupun cukup jarang untuk visit web, Kelompok ini adalah kelompok yang cukup sering merespon campaign kita dan yang paling sering menggunakan promo dalam sebulannya dengan rata-rata penggunaan promo sebanyak 3 kali dalam sebulan
- Kelompok ini juga merupakan kelompok yang paling sering melakukan transaksi pada platform kita dibandingkan dengan kelompok lainnya.

4. **High Spender**:
- Kelompok ini adalah kelompok dengan jumlah user terbesar kedua sebanyak 619 orang yang di dominasi oleh older_adults (>55 tahun) dan middle_aged_adults (36-55 tahun), yang dominan telah belum/tidak menikah dan belum mempunyai anak
- Dari segi pendapatan dan pengeluaran, kelompok ini mempunyai pendapatan dan pengeluaran paling besar di setiap bulannya, yang masing-masing sebesar IDR 70 Juta untuk total pendapatan setahun, dan IDR 1.1 Juta untuk pengeluaran dalam setahun
- Cluster ini cukup banyak yang merupakan non-organic dengan merespon campaign, namun memiliki jumlah penggunaan promo yang paling sedikit dibandingkan dengan yang lainnya.
- Kelompok ini adalah kelompok yang mempunyai conversion rate terbesar untuk membeli produk kita, dan kita jangan sampai kehilangan mereka.

Recommendation and Potential Impact

Recommendation:

1. Tetap monitor transaksi dan retensi dari kelompok High Spender, Fokus untuk tingkatkan service agar kelompok kelompok ini tidak churn

2. Untuk kelompok Mid Spender dapat dilakukan analisis lebih lanjut bagaimana agar meningkatkan transaksinya dengan memberikan rekomendasi yang lebih personal, serta analisis lebih dalam bagaimana untuk optimasi promo pada segmen ini dan tetap berbelanja di platform kita

3. Untuk kelompok Low Spender dan Risk to Churn, juga dapat dilakukan analisis lebih lanjut bagaimana meningkatkan rasio konversi visit to transaction, Mereka mempunyai jumlah visit yang cukup tinggi tapi tidak melakukan transaksi. Hal ini dapat disebabkan oleh produk ataupun harga yang tidak cocok.

Potential Impact:

- Jika kita fokus untuk terus monitor kelompok High Spender, kita akan tetap mendapatkan potensial GMV sebesar IDR 691 Juta, sedangkan untuk kelompok Mid Spender sebesar IDR 387 Juta

- Jika kita dapat optimasi promo yang di spend untuk Mid Spender (dengan asumsi reduksi 50%) kita dapat melakukan reduksi cost seevsar IDR 50 Juta

Source Code dapat dilihat disini